

# Impact of using a privacy model on smart buildings data for CO<sub>2</sub> prediction

Marlon P. da Silva<sup>1</sup>, Henry C. Nunes<sup>1</sup>, Charles V. Neu<sup>2</sup>, Luana T. Thomas<sup>1</sup>,  
Avelino F. Zorzo<sup>1</sup>, and Charles Morisset<sup>2</sup>

<sup>1</sup> Polytechnic School PUCRS, Porto Alegre, Brazil

<sup>2</sup> School of Computing Newcastle University, Newcastle upon Tyne, UK

**Abstract.** There is a constant trade-off between the utility of the data collected and processed by the many systems forming the Internet of Things (IoT) revolution and the privacy concerns of the users living in the spaces hosting these sensors. Privacy models, such as the SITA (Spatial, Identity, Temporal, and Activity) model, can help address this trade-off. In this paper, we focus on the problem of  $CO_2$  prediction, which is crucial for health monitoring but can be used to monitor occupancy, which might reveal some private information. We apply a number of transformations on a real dataset from a Smart Building to simulate different SITA configurations on the collected data. We use the transformed data with multiple Machine Learning (ML) techniques to analyse the performance of the models to predict  $CO_2$  levels. Our results show that, for different algorithms, different SITA configurations do not make one algorithm perform better or worse than others, compared to the baseline data; also, in our experiments, the temporal dimension was particularly sensitive, with scores decreasing up to 18.9% between the original and the transformed data. The results can be useful to show the effect of different levels of data privacy on the data utility of IoT applications, and can also help to identify which parameters are more relevant for those systems so that higher privacy settings can be adopted while data utility is still preserved.

**Keywords:** Privacy · CO<sub>2</sub> Prediction · Smart Buildings · Sensors Data

## 1 Introduction

The impact of the quality of an indoor environment on the well-being of its occupants is a relatively well-studied problem [2]. More than 20 years ago, Redlich *et al.* noted the increase of the Sick-Building Syndrome (SBS), which includes “upper-respiratory irritative symptoms, headaches, fatigue, and rash” [22]. Although they deemed  $CO_2$  as “an unlikely cause of SBS”, a study widely covered in the general press, clearly indicates potential health risks associated with chronic exposure to environmentally relevant elevations in ambient  $CO_2$ , including “inflammation, reductions in higher-level cognitive abilities and impact on different body organs” [11]. There is, therefore, a clear need for precise and reactive monitoring of indoor  $CO_2$ , to detect and prevent dangerous situations.

On the one hand, smart buildings deploy IoT architecture usually including  $CO_2$  and temperature sensors [32], intended to be used on new services that can be provided, many supported by Machine Learning (ML) techniques. Those services are intended to automatise management and optimise user comfort, security, and safety, quite often with a focus on occupancy measurement [5, 16, 17]. The need for  $CO_2$  monitoring is likely to push an increasing deployment of such systems.

On the other hand, this monitoring faces an increasing privacy concern related to ambient infrastructures. Naieni *et al.* for instance showed that although roughly half of the participants in a survey were comfortable or very comfortable with the collection of presence and temperature data, people nevertheless favour data collection in which they cannot be identified immediately and do not want inferences to be made from otherwise anonymous data [21].

Lately, new legislation has been introduced to support and regulate personal data usage and people’s privacy preferences, e.g. GDPR (General Data Protection Regulation)<sup>3</sup> in the EU/UK and LGPD (Lei Geral de Proteção de Dados)<sup>4</sup> in Brazil. A common principle is that of *data minimisation*, which specifies that a system should not collect and process more data than needed for its purpose. There is also a clear concern that users must be involved in data collection and processing and their preferences must be considered. As a result, a data protection system must work out a difficult trade-off: minimise data collection to satisfy as much as possible user privacy preference while avoiding a loss to data utility, which might reduce the efficiency of processing, and as a consequence could impact the overall utility of the data to the provided services.

In this paper, we explore this trade-off in the context of a real-world smart building by evaluating how different ML methods perform to predict  $CO_2$  when different privacy levels are defined. We also evaluate how different levels of privacy impact data utility in comparison to when the whole data is available.

The remainder of this work is organised as follows. Section 2 presents background on smart buildings, data privacy, and machine learning. Section 3 presents our methodology to configure the dataset and to build our implementation. An experimental evaluation and discussion are presented in Section 4. Section 5 discusses recent works on privacy in smart spaces and  $CO_2$  prediction using smart buildings sensor’s data. Finally, Section 6 concludes this work and indicates some future work directions.

## 2 Background

With the revolution of IoT equipment, many smart buildings are emerging, especially in universities and business offices. They are responsible for collecting a huge amount of data from many people every day. In light of this, there is a growing concern about the privacy of these data.

<sup>3</sup> <https://gdpr.eu/>

<sup>4</sup> <https://www.serpro.gov.br/lgpd/menu/a-lgpd/o-que-muda-com-a-lgpd>

For our research, we collected data from a smart building located at a university in England, applied different settings of the SITA privacy model that will be discussed in the next sections, and used machine learning algorithms to check if there were changes regarding the usability of the data.

Thus, we investigated and collected some information on the topics covered, which are organised as follows in the next sections: Section 2.1 deals with smart buildings, Section 2.2 presents the SITA model, and Section 2.3 describes some machine learning algorithms applied in this work.

## 2.1 Smart buildings sensors and data

Smart building is a term that has its origins in the larger scenario of building automation, which is the set of practices aimed to improve the control of a building by electronic means. From building automation, emerges the concept of intelligent building, when, in addition to control, we also have historical data enabling us to make predictions [4].

The drivers for the development of buildings can be said to revolve around adding value to a building [27]. Reducing energy consumption has now become a driver in its own right, due to increasingly stringent regulations and awareness of climate change. This is recognised in modern buildings as a significant design criterion [9]. In order to achieve these requisites, there are four specific approaches to follow:

- the methods by which building operation information is gathered and responded to (intelligence);
- the interaction between the occupants and the building (control);
- the building’s physical form (materials and construction); and
- the methods by which building use information is collected and used to improve occupant performance (enterprise).

Smart Buildings are Intelligent Buildings, but with additional, integrated aspects of adaptable control, enterprise and materials, and construction. In Smart Buildings, the four methods used to meet the drivers to building progression, mentioned previously, are developed alongside each other, utilising information from one in the operation of another. This is in contrast to Intelligent Buildings, which have largely developed intelligence independently of the other methods.

## 2.2 Privacy Model

In recent years there was an increase in demand for privacy techniques [25] [24]. This comes in line with an increase in awareness of society to how easily data is collected, distributed, and used in the information age. One consequence is the creation of legislation in different jurisdictions that address this topic, GDPR, LGPD, and PDP<sup>5</sup> are a few examples. This legislation tries to organise how

<sup>5</sup> <https://prsindia.org/billtrack/the-personal-data-protection-bill-2019>

data is treated and protected. To tackle the privacy problem the SITA model was developed.

SITA [1] is a conceptual model that empowers end-users with the ability to control their privacy. It is based on a granular approach to control privacy in applications, and also, uses Maeda's ten laws of simplicity [18] as a design philosophy. As result, an end-user has a simple and intuitive method of controlling how an application can distribute its data.

The embedded privacy control in most applications works binary, where a user can block the application from sharing all his data or allow it to share all his data in the application. SITA proposes the use of different levels as a way to remediate this. The user can control how much information he is sharing, this, however, comes with a cost. Less information that is shared in the application can degrade an application service because of the lack of precise information. As consequence, a user will need to set the application privacy control in some way so that the application is usable for his needs and also does not share so much information. This trade-off is very common in privacy applications, known as the **privacy-utility trade-off**. Other frameworks work on a similar premise, allowing more control over the privacy settings [23] [10], however, these models are in general complex, which hinders their widespread usage.

The model is divided into four dimensions: Spatial, referring to the user's location data, such as GPS position, address, and others; Identity, related to the user's personal identification data, such as ID, Name, and Gender; Temporal, date and time information about user activity in the application; Activity, sensitive data about user behaviour, situational data, and preferences. All the data shared with the application developer can be categorised into one of these four groups.

Each dimension can be assigned a level from zero to four. The level represents the amount of privacy for that specific dimension. Where zero represents no access to the data and complete privacy. On the other extreme four represents full access and no privacy protection. The values in-between allow controlling the shared data using aggregation, and obfuscation techniques to granularly control privacy. In that case, the amount of information shared is something between no shared data at all (level zero), and total access to the data (level four). These in-between values are application specific and need to be created by the developer.

An end user can set the level for each dimension, which is called a **SITA configuration**. The resulting data shared will apply the privacy level for each dimension and share the data with the application. A user can for instance set The SITA level as Spatial two, Identity, three, Temporal zero, and Activity four. The resulting shared data would include Spatial and Identity data modified by some anonymization aggregation, and obfuscation techniques, all the temporal data, and no activity data. In this work to identify different configurations a sequence of four numbers is used. Each number represents the value for one dimension. For example, 4343 means  $S = 4$ ,  $I = 3$ ,  $T = 4$ , and  $A = 3$ .

### 2.3 Supervised machine learning algorithms

Supervised machine learning algorithms use previously labelled data to train machine learning models. One of its main uses is to allow the inference of further data to a label. There are several algorithms, and variants, that use this method [12]. For this work, we describe the algorithms used in our experiment.

**Linear regression (LR)** This method is used to predict the value of a variable, named the dependent variable, and based on the value of other variables, named the independent variables. The independent variables are used to compose a linear equation. The dependent variable of a new data entry is predicted using the value of the independent variables as input in the linear equation, resulting in the predicted value.

**Ridge Regression (RR)** Multiple regression models, similar to linear regression, create an equation using independent variables to predict a dependent variable. However, unlike linear regression, it creates an equation composed of multiple coefficients. One problem that such an approach suffers is that highly correlated independent variables degrade the model performance. Ridge Regression solves this problem by substituting the least square estimators from multiple-regression models with a ridge regression estimator.

**Random forest (RF)** Random forest uses multiple decision trees to output a label for a new data entry. The label which is outputted more by the decision trees is the one attributed to the new data entry. Decision trees are multiple decision points in a tree-like format. Each non-leaf node contains a decision that will induce entry to a leaf node. The leaf node contains the label that the data entry will assume. The multiple trees in the random forest are created with some degree of variance. This method reduces the bias and overfitting that using a single decision tree can result in.

**Gradient Boosting Regressor (GBR)** This model is a variant of the ensemble methods, like RF. In these methods, multiple simple models are combined into a more complex and precise one. For GBR the simpler models usually are decision trees. The result is a more robust model overall. This approach can find any non-linear relationship in the dataset and can treat missing values, and outliers. The main difference between RF and GBR is the process of creating the decision tree and combining the results.

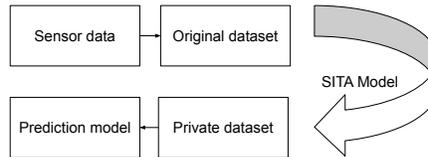
**Decision Tree Regressor (DTR)** A decision tree is the smaller model that is used in the RF and GBR. It is a classifier in the form of a tree. Each node is a decision and the leaves are the labelled value. A new entry starting at the root is directed at each node to children nodes based on its value. At the leaf node, it receives its label. The tree is built by partitioning the training dataset and building a decision node based on the partitioning.

### 3 Experiment

For our experiments, we collected data from a real scenario, the Urban Sciences Building(USB) at Newcastle<sup>6</sup>. In the next sections, we describe the methodology used for collecting and pre-processing the data, applying the privacy model, and the metrics used to measure the performance of the machine learning algorithms.

#### 3.1 Privacy model for IoT datasets in $CO_2$ prediction

In IoT scenarios, data from sensors can be stored in datasets. A dataset can be useful for history, use in prediction models and pattern detection, forensics, and other cases. However, the dataset can also be used in a malicious way. For example, a malicious data holder can use the data as part of a linkage attack to infringe upon individual privacy. To mitigate this risk one technique is to reduce the precision of data entries in the dataset. This diminishes the data precision, hindering malicious privacy attacks. This approach is used by the SITA model, from an original dataset it generates a private dataset where the original data entries are less precise.



**Fig. 1.** Proposed privacy model used for  $CO_2$  prediction

Figure 1 presents how we apply the SITA model to add different privacy levels to the original data. First, the data is collected from multiple sensors and aggregated in the original dataset. This data will suffer a SITA transformation based on a SITA Configuration. It is important to note that the same configuration is applied to all the datasets, instead of a configuration for a specific user’s data or data entries. The reason for this is that we intend to isolate and analyse the impact of the transformation for each SITA dimension on a machine learning  $CO_2$  prediction model. The result of the SITA Transformation is a private dataset, a dataset with increased security against privacy attacks. This private dataset is then used to create a prediction model using machine learning techniques.

This approach reflects the possibility that an IoT environment can automatically transform an original dataset, or data as it is added to a dataset using the SITA model. The resulting private dataset can be the only information made

<sup>6</sup> <https://api.usb.urbanobservatory.ac.uk>

available to the data holder. This guarantees an increase in privacy. However, this comes with the cost of data utility, since the modification to the original dataset will decrease its data utility and prejudice the prediction models created from it.

### 3.2 Attack model

Here we present a simple model that can be used to exploit the CO<sub>2</sub> readings from a sensor, or its prediction if accurate, to determine which specific person is in a room. There is a number of studies that suggest that people who weigh more produce more CO<sub>2</sub> [19], and there are studies that suggest that males produce more CO<sub>2</sub> than females [35]. Also, there is the ASHRAE Standard 62.1 [3] used to predict CO<sub>2</sub> emission inside a building and control air quality, it receives as input the metabolic rate, which is highly based on a person's body composition (fat, muscle, bones, and etc.). In our model, the potential difference of CO<sub>2</sub> emission between two individuals will change the CO<sub>2</sub> readings of the room, allowing it to be used to identify who is inside a room.

The scenario for this model is a small closed room that is used by just two people. These two people have a significant difference in body composition and are of different sex. We will identify them as Alice (50kg) and Bob (90kg), a third person, Eve, wants to identify who is inside the room without their consent. Eve has some background information, she knows the sex of Alice and Bob and their approximate weight. Eve also has access to the CO<sub>2</sub> readings of the room.

In Figure 2 we summarise our model. The CO<sub>2</sub> readings of the room (a), and the background information (b) will be the input of the model of the gas dispersion model (c), that model will as result identify who is inside the room. The gas dispersion model can use the ASHRAE Standard 62.1 [3] to predict the CO<sub>2</sub> present in the room when it is empty, Alice is in it, Bob is in it, or both. Using this prediction and the actual CO<sub>2</sub> readings Eve can then predict who is in the room. The gas dispersion model is not the only option viable for Eve, if she has historic data of the room CO<sub>2</sub> she can use an unsupervised machine learning algorithm to cluster the readings in four groups, group 1 when the room is empty, group 2 when Alice is in the room, group 3 when Bob is in the room, and group 4 when both are in the room. Finally, if Eve has a history of CO<sub>2</sub> readings and who is present in the room even other supervised machine learning algorithms are viable.

This model is very simplistic and in more complex scenarios can be ineffective. A few possible changes that can impact this model include: Changing the number of people that use the room, this would increase the complexity of isolating the CO<sub>2</sub> emission of one person, and more people more complex. People with very similar profiles, same-sex same weight, or very close. A big room would disperse the emitted CO<sub>2</sub>, and the presence of people inside the room would change very little in the room readings, making it impossible to distinguish who is inside. Ventilation can also disperse the CO<sub>2</sub>, making it impossible to distinguish who is inside.

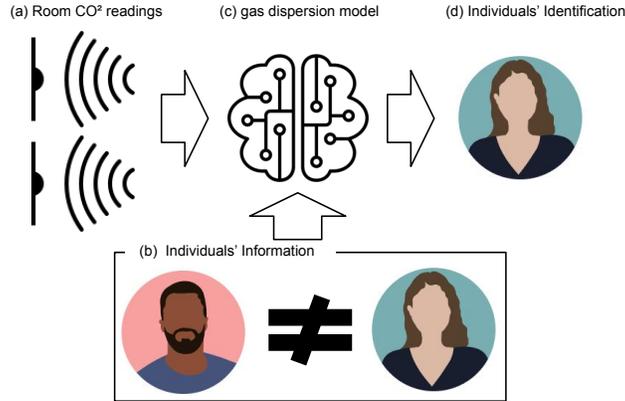


Fig. 2.  $CO_2$  individuals' privacy attack model

### 3.3 Description of the Experiment

The objective of our experiment is to analyse the impact of the SITA privacy model in  $CO_2$  predictions that use machine learning. More specifically the experiment aims to analyse how changing one dimension of the SITA configuration impacts the performance of the prediction model. Privacy techniques, such as SITA, suffer from the privacy-utility trade-off, increased privacy decreases the data utility which will impact the prediction model.

We briefly summarise the experiment conducted in the following sequence of steps and in Figure 3. More detail will be provided in the next subsections:

- **Collecting original dataset:** The first step was the collection of the dataset for the prediction model. We used data publicly available from the USB.
- **Data Transformation:** The original dataset was transformed to better suit the SITA transformation.
- **SITA Transformation:** Different datasets were created from different SITA configurations. We aim to analyze the isolated impact of each SITA dimension. Thus, for each dimension, we changed its level from 0 to 4 while keeping the others in a fixed state. We will analyse the SITA configurations X444, 44X4, and 444X, where X is a number between 0 and 4. This transformation results in five different private datasets for each dimension.
- **ML Training:** Our prediction models are created using LR, RR, RF, GBR, and DTR. The selected techniques are based on the work of Wibisono *et al.* [33]. Each technique is used with all datasets generated from the different SITA configurations.
- **Analysis:** To analyse the impact of a SITA configuration in the prediction model we utilise common ML metrics. The metrics are  $R^2$  score, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The chosen metrics are also based on the work of Wibisono *et al.* [33].

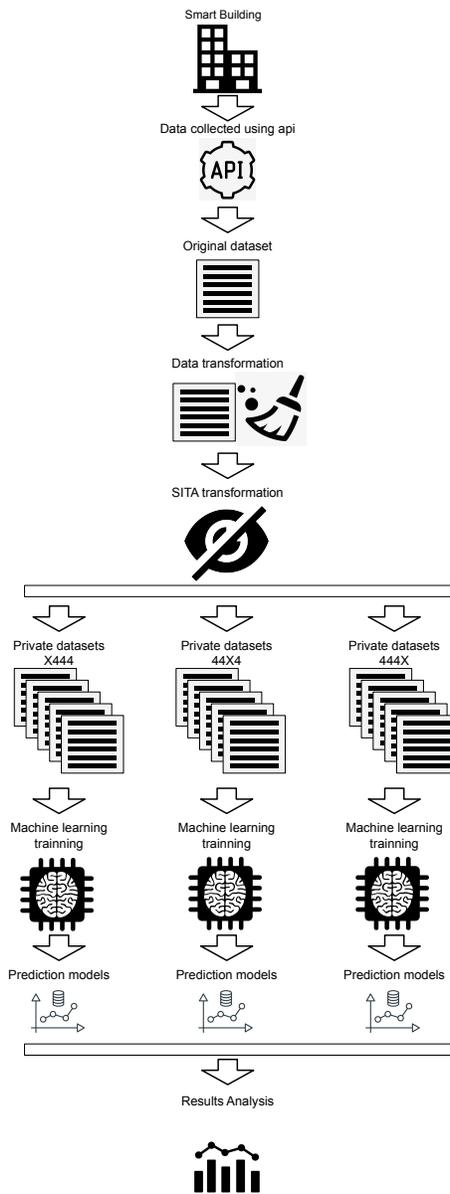


Fig. 3. Experiment summary

### 3.4 Dataset Scenario and data collection

Our scenario is composed of the USB. It is the largest open platform for urban sensing data in Newcastle. The building includes multiple IoT sensors that can be freely accessed online through their website. An API is also provided, which allows other applications to access the data, in real-time, and also previous data.

The dataset was extracted using a script developed by our team, collecting data from October 2018 to March 2020. We collected data from five sensors: humidity, temperature, occupancy, brightness, and  $CO_2$ . The data is also organised by rooms, with the readings of the sensors in each of these rooms. The rooms differ in size, sensors available, and usage. The historical data is stored in the API by sensors, in different JSON files. Therefore, we had to consolidate all the data into a single file and remove all records that contained at least one missing data.

### 3.5 Data Transformation

To remove all the outliers from our dataset, we set a range of values for each feature. So we have  $CO_2$  values ranging from 0 ppm to 1,000 ppm (ASHRAE limit for healthy environments<sup>7</sup>) and temperature values ranging from 0°C to 50°C. The relative humidity values ranged from 0% to 100% and the brightness values ranged from 0 lm to 2000 lm.

After completing the data cleaning as described above, we have a new dataset with about 200,000 records. This dataset is ready to be used in the SITA transformation and in future works.

### 3.6 SITA Data Transformation

Each level of the SITA parameters corresponds to a specific operation on those variables related to that parameters. Below we describe those relationships, and the transformations performed at each level. The Identity dimension is absent in our work, since there is no individual data stored in the datasets, and because of that its respective operations are disabled.

The Spatial dimension is represented by data regarding the room and the zone of each entry. Given a sample input  $G.024, 2$ , the operations follow:

- Level 0: all data deleted. Output: *deleted,deleted*
- Level 1: only the general location is given. Output: *building,deleted*
- Level 2: only the ground of each room is given. Output: *Ground Floor,deleted*
- Level 3: returns full information about the room, omitting the zone data. Output: *G.024, deleted*
- Level 4: no transformations are applied. Output: *G.024, 2*

For the Temporal dimension, we took as input the datetime parameter. We present the results given a sample input  $20181011141735$ .

<sup>7</sup> <https://www.ashrae.org/about/position-documents>

- Level 0: all data deleted. Output: *deleted,deleted*
- Level 1: year and month, with day fixed at 01. Output: *20181001,deleted*
- Level 2: date. Output: *20181011,deleted*
- Level 3: date and hour. Output: *20181011, 140000*
- Level 4: no transformations are applied. Output: *20181011, 142735*

For the Activity dimension, we considered the following attributes: CO2, Temperature, Humidity, and Brightness. For the sample input *287.0,27.6,63.8,25.0*, we have the following operations:

- Level 0: all data deleted. Output: *deleted,deleted,deleted,deleted*
- Level 1: values are rounded up to the two rightmost digits. Output: *300,0,100,0*
- Level 2: values are rounded up to the rightmost digit. Output: *290,30,60,30*
- Level 3: decimal digits are removed. Output: *287,27,63,25*
- Level 4: no transformations are applied. Output: *287.0,27.6,63.8,25.0*

For the experiment, a SITA configuration is applied to all the entries in the dataset. We applied the following configurations to the original dataset: 4444, 3444, 2444, 1444, 0444, 4434, 4424, 4414, 4404, 4443, 4442, 4441, and 4440. To better organise we will refer to a group of operations that alter the same dimension using X for the dimension it is altering. For example, X444 refers to configurations 4444, 3444, 2444, 1444, and 0444. Note that this transformation from the original dataset resulted in multiple private datasets, one new dataset from each configuration.

### 3.7 Machine Learning training

The ML models were trained using the Kaggle<sup>8</sup> platform, in a remote computing environment with 4 CPUs and 16 Gigabytes of RAM. The library used for the training was the scikit-learn<sup>9</sup> version 1.0.2. Before the training, since the algorithms here studied only work with numerical data, we transformed all textual data into numerical over each dataset. After this, the datasets were split into training/testing in a proportion of 80/20 utilising random sampling, with a random state of 10. Over these, we applied the *KFold()* method from scikit-learn, with ten splits and setting the parameter *shuffle* to *true*, to avoid overfitting the model. Each regressor method was then instantiated using their default implementations; the  $R^2$ , MAE, and RMSE scores were calculated using the function *cross\_val\_score()*, indicating in the *score* parameter the respective metric.

## 4 Results and Discussion

This section summarises the results of our experiments. We have three sets of SITA configurations, each one varying by one dimension and keeping the other

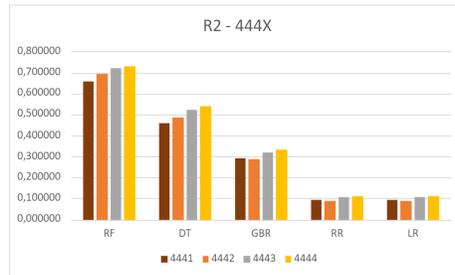
<sup>8</sup> <https://www.kaggle.com>

<sup>9</sup> <https://scikit-learn.org>

three unmodified (i.e. level 4 of the model). We did this to better comprehend the impact of applying different SITA options over the dataset.

We measured the results accordingly to three metrics:  $R^2$  score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).  $R^2$  score, also known as the Coefficient of Determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s). In other words, it represents the correlation between the predicted outcomes of a model and their real values [7]. Mean Absolute Error is the average difference between predicted and real values, representing how much the model misses the expected value. The Root-mean-square error, like MAE, is also an error metric, but here the measurement is calculated by the square root of the average of squared differences between prediction and actual observation. Although similar (with MAE being usually recommended, due to its easier interpretation [34]), MAE and RMSE exhibit distinct behaviours in specific circumstances, such as with large test sizes; thus, a combination of both metrics are often required to assess model performance [6].

For each SITA configuration analysed, we applied ten-fold cross-validation over the trained models, and collected their  $R^2$ , MAE, and RMSE values at the end of each execution; after all executions were completed, we calculated the average scores of each model/configuration pair.



**Fig. 4.**  $R^2$  score for machine learning algorithms in 444X SITA dimension.

For every specific setting, we ran all five algorithms, and measured the results according to three parameters: the coefficient of determination ( $R^2$  score - figures 2 to 4), the Mean Absolute Error (MAE - figures 5 to 7), and the Root Mean Squared Error (RMSE - figures 8 to 10). Since there are significant performance differences between Random Forest and Decision Tree models compared to the other algorithms evaluated, we will focus our analysis on these first two.

Regarding the  $R^2$  score, the Random Forest algorithm outperforms all other approaches analysed, with an average value of 74.29% for the baseline data. When taking the Activity dimension as variable, the minimum average score of this model is 66.05%. This represents a performance decrease of 9.44% regarding to the baseline. Even in this worst-case scenario, the Random Forest model still



Fig. 5.  $R^2$  score for machine learning algorithms in X444 SITA dimension.

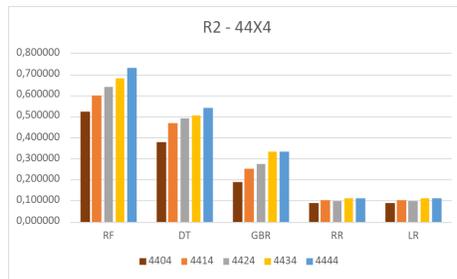


Fig. 6.  $R^2$  score for machine learning algorithms in 44X4 SITA dimension.

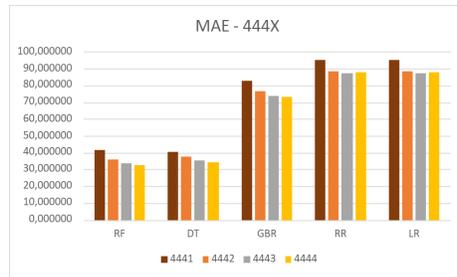


Fig. 7. MAE score for machine learning algorithms in 444X SITA dimension.

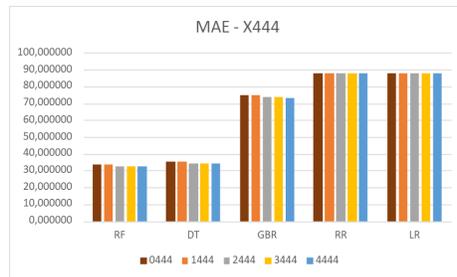


Fig. 8. MAE score for machine learning algorithms in X444 SITA dimension.

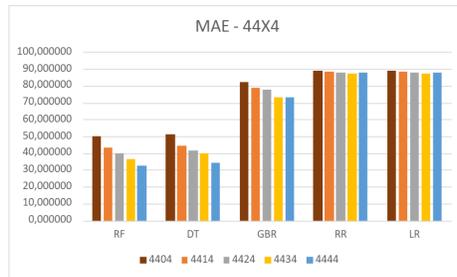


Fig. 9. MAE score for machine learning algorithms in 44X4 SITA dimension.

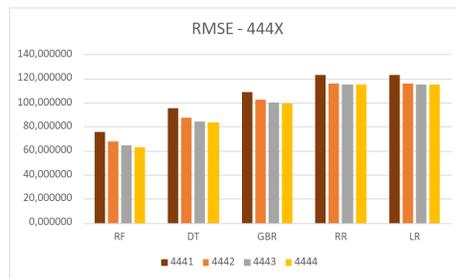


Fig. 10. RMSE score for machine learning algorithms in 444X SITA dimension.



Fig. 11. RMSE score for machine learning algorithms in X444 SITA dimension.

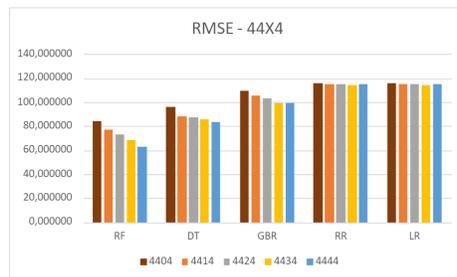


Fig. 12. RMSE score for machine learning algorithms in 44X4 SITA dimension.

outperforms the second-best algorithm (Decision Tree) by 23.62%, even when considering the baseline for the latter. It is important to observe that we did not apply level 0 of SITA transformations in that dimension since this operation would erase all CO2 data, thus making it impossible to predict its levels.

For the Temporal dimension, the minimum average score of Random Forest is 51.80% for the Temporal privacy setting of 0 (i.e. all data deleted). This represents a performance decrease of 28.98% regarding the baseline. In this worst-case scenario for the Random Forest model, the Decision Tree baseline performance is 3.05% superior, with all other temporal settings of the Random Forest performing better than the Decision Tree baseline, and all other temporal settings of the Decision Tree underperforming the Random Forest worst-case.

When considering the Spatial dimension, the Random forest model produces an average  $R^2$  score of 71.86% when the privacy setting is at 0, a decrease of 1.73%. This value is 33.00% higher than the Decision Tree score for the baseline case, having this model in the same privacy level a score of 51.16%, being 28.81% lower than RF.

Analysing the Mean Absolute Error, the lowest score for the baseline was obtained running the Random Forest model, with a result of 32,40%. With the Activity dimension as variable, the maximum average score of such algorithm equals to 41,28% when privacy level = 1 for Activity. This represents an error increase of 27,38% regarding to the baseline. The Decision Tree algorithm produces an MAE value of 40,60%, which represents a decrease of 1,64% in relation to the Random Forest.

For the Temporal dimension, we have a maximum average score of 49,85% (privacy level = 0), 53,50% higher than the baseline. The same configuration when applied to Decision Tree produces an MAE score of 50,93%, 2,17% higher than the Random Forest model.

The Spatial dimension has an MAE score of 33,63% on the strictest privacy level for the RF model. This represents an increase of 3,56%, considering the baseline. The Decision Tree model, conversely, has an MAE score of 35,49%, an increase of 5,53% from the RF.

Finally, we look at the RMSE metric. The lowest value was obtained with the Random Forest model, with a score of 62.99%. For the Activity dimension, the highest value was 75.47%, 13.83% higher than the baseline. When we set Temporal and Spatial domains as variables, their respective scores were 84.20% (an increase of 33.67%) and 64.31% (2.09% higher than the baseline).

#### 4.1 Discussion

Firstly, our results confirm the experimental data presented in [33], regarding the performance of the ML algorithms then analysed. Linear Regression and Ridge Regression, being fairly simple algorithms, are expected to perform poorly than more sophisticated methods, especially on large datasets. The other algorithms used in our experiments can be seen as belonging to the same family, with Decision Tree being the basis for both Random Forest and Gradient Boosting. However, some careful tuning is necessary for the latter to achieve good results,

which makes it harder to apply the method over different domains. Regarding Decision Tree and Random Forest methods, since the latter is an averaging of multiple instances of the first (thus mitigating possible errors due to overfitting), the results of our experiments confirm the expected performances. We also show that RF is the only algorithm between those analysed that produces  $R^2$  scores over 70%. When analyzing the Mean Absolute Error, our results show that the performances of RF and DT algorithms are very close, but the RMSE values present a more significant difference between these two methods. This can be explained by the fact that RMSE has a tendency to be increasingly larger than MAE as the test sample size increases, thus exacerbating small differences in MAE values between the two approaches.

Another discussion can be made about the impact of applying our SITA implementation over the chosen dataset. Our results show that the dimensions present different sensibilities to more restrictive privacy settings. Taking the  $R^2$  score, we show that the Spatial dimension is the least affected, and the Temporal dimension the most affected, with Activity being in an intermediate place. This can be used to better understand the importance of different variables in applying ML techniques. Also, by analysing the scores of each privacy setting, we observe that the Activity dimension has a score below 70% when the privacy setting is lower than 3; the same occurs for the Temporal dimension with privacy setting lower than 4 (reflecting the higher sensitivity of this dimension), and it is not observed in the Spatial dimension in any configuration. With this we demonstrate that it is possible, through different SITA settings, to improve the users' privacy and keep ML services functional.

An interesting approach for further research on this topic is the use of other machine learning algorithms, including more powerful techniques such as deep learning. Exploring other domains such as healthcare, social media, and other IoT scenarios for example are also interesting further directions.

## 5 Related Work

There are numerous works related to the prediction of  $CO_2$  in IoT environments using machine learning algorithms. The  $CO_2$  monitoring is an important component of controlling the air quality of a room, which when correctly managed provides well-being, controls general air pollution, and detects potential harms, such as fire. Creating a prediction model can be positive in cases presented by Kapoor *et al.* [14] where smart sensors are not available, also a model can be used to help in the building design. In his work, they present a model working with multiple machine learning algorithms and achieve a precise model.

Other works are developed in a similar fashion using machine learning algorithms in an IoT scenario to create a prediction model for  $CO_2$ . Vanus *et al.* [31] use the value of other sensors like temperature and humidity to predict the  $CO_2$  value in a Smart Home scenario. In another study, Sharma *et al.* [26] describe the building of a sensor network to detect different pollutant gases beyond  $CO_2$ , although still in development a model is described. An artificial neural network

is used to predict air quality and to fully control an IoT network, including air-conditioning, and ventilation based on the work of Tagliabue *et al.* [29].

There are other works in the prediction of air quality and CO<sub>2</sub> [30] [13] [15]. However, the use of such prediction models and privacy models as the one presented in Section 2.2 are not common. The readings of a CO<sub>2</sub> with other background information can be used in a linkage attack [20] in the original dataset, or real-time data, to discover information about individuals, for example, who was present in a room, patterns of movements in a building, among others.

## 6 Final consideration and future work

In this work, we analysed the trade-off between privacy and utility for CO<sub>2</sub> prediction on a real dataset in the context of smart buildings. Therefore, several transformations were implemented on the original data to simulate different privacy levels and generate new transformed datasets that were used as input to train five distinct machine learning models for CO<sub>2</sub> prediction.

The results show that the performance of Regression based machine learning techniques is lower than decision Tree-based techniques. The use of the privacy model, as expected, deteriorated the performance of all algorithms. More aggressive SITA configurations resulted in worse performance and each dimension has a different impact on the prediction models. The highest impact was observed when higher privacy levels were simulated on the Temporal dimension.

As future research directions, our model could be improved by using **Syntactic Anonymity** [28] with SITA to increase even more the data privacy. To the best of our knowledge, there is no work of this kind yet. Also, the inclusion of **Differential Privacy** [8] is another possibility that could improve even more the privacy model, since it is a more powerful privacy definition than syntactic anonymity.

## References

1. Andersen, M.S., Kjargaard, M.B., Grønbæk, K.: The sita principle for location privacy — conceptual model and architecture. In: 2013 International Conference on Privacy and Security in Mobile Systems (PRISMS). pp. 1–8 (2013). <https://doi.org/10.1109/PRISMS.2013.6927184>
2. Arif, M., Katafygiotou, M., Mazroei, A., Kaushik, A., Elsarrag, E., et al.: Impact of indoor environmental quality on occupant well-being and comfort: A review of the literature. *International Journal of Sustainable Built Environment* **5**(1), 1–11 (2016)
3. Ashrae, A., Standard, A.: 62.1. 2007, ventilation for acceptable indoor air quality. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA (2007)
4. Buckman, A.H., Mayfield, M., Beck, S.B.: What is a smart building? *Smart and Sustainable Built Environment* (2014)
5. Candanedo, L.M., Feldheim, V.: Accurate occupancy detection of an office room from light, temperature, humidity and co<sub>2</sub> measurements using statistical learning models. *Energy and Buildings* **112**, 28–39 (2016)

6. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature. *Geoscientific model development* **7**(3), 1247–1250 (2014)
7. Draper, N.R., Smith, H.: *Applied regression analysis*, vol. 326. John Wiley e Sons (1998)
8. Dwork, C.: Differential privacy: A survey of results. In: *Proceedings of the 5th International Conference conference on theory and applications of models of computation (TAMC)*. pp. 1–19 (2008)
9. GhaffarianHoseini, A., Dahlan, N.D., Berardi, U., GhaffarianHoseini, A., Makaremi, N., GhaffarianHoseini, M.: Sustainable energy performances of green buildings: A review of current theories, implementations and challenges. *Renewable and Sustainable Energy Reviews* **25**, 1–17 (2013)
10. Gupta, A., Kalra, A., Boston, D., Borcea, C.: Mobisoc: a middleware for mobile social computing applications. *Mobile Networks and Applications* **14**(1), 35–52 (Feb 2009). <https://doi.org/10.1007/s11036-008-0114-9>, <https://doi.org/10.1007/s11036-008-0114-9>
11. Jacobson, T.A., Kler, J.S., Hernke, M.T., Braun, R.K., Meyer, K.C., Funk, W.E.: Direct human health risks of increased atmospheric carbon dioxide. *Nature Sustainability* **2**(8), 691–701 (2019)
12. Jiang, T., Gradus, J.L., Rosellini, A.J.: Supervised machine learning: A brief primer. *Behavior Therapy* **51**(5), 675–687 (2020). <https://doi.org/https://doi.org/10.1016/j.beth.2020.05.002>, <https://www.sciencedirect.com/science/article/pii/S0005789420300678>
13. Kadam, P., Vijayumar, S.: Prediction model: Co2 emission using machine learning. In: *2018 3rd International Conference for Convergence in Technology (I2CT)*. pp. 1–3 (2018). <https://doi.org/10.1109/I2CT.2018.8529498>
14. Kapoor, N.R., Kumar, A., Kumar, A., Kumar, A., Mohammed, M.A., Kumar, K., Kadry, S., Lim, S.: Machine learning-based co2 prediction for office room: A pilot study. *Wireless Communications and Mobile Computing* (Mar 2022)
15. Khorram, M., Faria, P., Abrishambaf, O., Vale, Z., Soares, J.: Co2 concentration forecasting in an office using artificial neural network. In: *2019 20th International Conference on Intelligent System Application to Power Systems (ISAP)*. pp. 1–6 (2019). <https://doi.org/10.1109/ISAP48318.2019.9065944>
16. Labeodan, T., Zeiler, W., Boxem, G., Zhao, Y.: Occupancy measurement in commercial office buildings for demand-driven control applications—a survey and detection system evaluation. *Energy and Buildings* **93**, 303–314 (2015)
17. Lam, K.P., Höynck, M., Dong, B., Andrews, B., Chiou, Y.S., Zhang, R., Benitez, D., Choi, J., et al.: Occupancy detection through an extensive environmental sensor network in an open-plan office building. *IBPSA Building Simulation* **145**, 1452–1459 (2009)
18. Maeda, J.: *The Laws of Simplicity. Simplicity: Design, Technology, Business, Life*, MIT Press (2006), [https://books.google.com.br/books?id=1h9gRQ4i\\_zwC](https://books.google.com.br/books?id=1h9gRQ4i_zwC)
19. Magkos, F., Tetens, I., Bügel, S.G., Felby, C., Schacht, S.R., Hill, J.O., Ravussin, E., Astrup, A.: The environmental foodprint of obesity. *Obesity* **28**(1), 73–79 (Dec 2019). <https://doi.org/10.1002/oby.22657>, <https://doi.org/10.1002/oby.22657>
20. Merener, M.M.: Theoretical results on de-anonymization via linkage attacks. *Trans. Data Privacy* **5**(2), 377–402 (aug 2012)
21. Naeini, P.E., Bhagavatula, S., Habib, H., Degeling, M., Bauer, L., Cranor, L.F., Sadeh, N.: Privacy expectations and preferences in an {IoT} world. In: *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. pp. 399–412 (2017)

22. Redlich, C.A., Sparer, J., Cullen, M.R.: Sick-building syndrome. *The Lancet* **349**(9057), 1013–1016 (1997)
23. Roman, M., Hess, C., Cerqueira, R., Ranganathan, A., Campbell, R., Nahrstedt, K.: A middleware infrastructure for active spaces. *IEEE Pervasive Computing* **1**(4), 74–83 (2002). <https://doi.org/10.1109/MPRV.2002.1158281>
24. Sarwar, K., Yongchareon, S., Yu, J., Ur Rehman, S.: A survey on privacy preservation in fog-enabled internet of things. *ACM Comput. Surv.* **55**(1) (nov 2021). <https://doi.org/10.1145/3474554>, <https://doi.org/10.1145/3474554>
25. Seliem, M., Elgazzar, K., Khalil, K.: Towards privacy preserving iot environments: A survey. *Wireless Communications and Mobile Computing* **2018**, 1032761 (Nov 2018). <https://doi.org/10.1155/2018/1032761>, <https://doi.org/10.1155/2018/1032761>
26. Sharma, P.K., De, T., Saha, S.: Iot based indoor environment data modelling and prediction. In: 2018 10th International Conference on Communication Systems & Networks (COMSNETS). pp. 537–539 (2018). <https://doi.org/10.1109/COMSNETS.2018.8328266>
27. Smith, S.: Intelligent buildings. In: Design and Construction, pp. 60–82. Routledge (2007)
28. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(05), 571–588 (2002). <https://doi.org/10.1142/S021848850200165X>
29. Tagliabue, L.C., Re Cecconi, F., Rinaldi, S., Ciribini, A.L.C.: Data driven indoor air quality prediction in educational facilities based on iot network. *Energy and Buildings* **236**, 110782 (2021). <https://doi.org/https://doi.org/10.1016/j.enbuild.2021.110782>, <https://www.sciencedirect.com/science/article/pii/S0378778821000669>
30. Vanus, J., Krestanova, A., Kubicek, J., Gorjani, O., Penhaker, M., Oczka, D.: Using wavelet transformation for prediction co<sub>2</sub> in smart home care within iot for monitor activities of daily living. In: Nguyen, N.T., Chbeir, R., Exposito, E., Anioté, P., Trawiński, B. (eds.) *Computational Collective Intelligence*. pp. 500–509. Springer International Publishing, Cham (2019)
31. Vanus, J., Martinek, R., Bilik, P., Zídek, J., Dohnalek, P., Gajdos, P.: New method for accurate prediction of co<sub>2</sub> in the smart home. In: 2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings. pp. 1–5 (2016). <https://doi.org/10.1109/I2MTC.2016.7520562>
32. Verma, A., Prakash, S., Srivastava, V., Kumar, A., Mukhopadhyay, S.C.: Sensing, controlling, and iot infrastructure in smart building: A review. *IEEE Sensors Journal* **19**(20), 9036–9046 (2019). <https://doi.org/10.1109/JSEN.2019.2922409>
33. Wibisono, A., Wisesa, H.A., Habibie, N., Arshad, A., Murdha, A., Jatmiko, W., Gamal, A., Hermawan, I., Aminah, S.: Dataset of short-term prediction of co<sub>2</sub> concentration based on a wireless sensor network. *Data in Brief* **31**, 105924 (2020). <https://doi.org/https://doi.org/10.1016/j.dib.2020.105924>
34. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research* **30**(1), 79–82 (2005)
35. Yang, L., Wang, X., Li, M., Zhou, X., Liu, S., Zhang, H., Arens, E., Zhai, Y.: Carbon dioxide generation rates of different age and gender under various activity levels. *Building and Environment* **186**, 107317 (2020). <https://doi.org/https://doi.org/10.1016/j.buildenv.2020.107317>, <https://www.sciencedirect.com/science/article/pii/S0360132320306880>